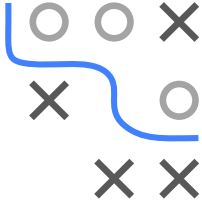
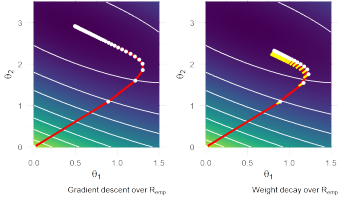
Introduction to Machine Learning



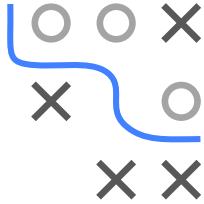
Regularization
Weight Decay and L2



Learning goals

- Understand why *L2* regularization in combination with gradient descent is equivalent to weight decay
- Understand how weight decay changes the optimization trajectory

WEIGHT DECAY VS. L2 REGULARIZATION



Let us optimize the *L2*-regularized risk of a model $f(\mathbf{x} \mid \boldsymbol{\theta})$

$$\min_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$$

by gradient descent. The gradient is

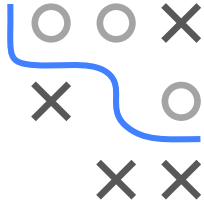
$$\nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \boldsymbol{\theta}.$$

We iteratively update $\boldsymbol{\theta}$ by step size α times the negative gradient

$$\begin{aligned} \boldsymbol{\theta}^{[\text{new}]} &= \boldsymbol{\theta}^{[\text{old}]} - \alpha \left(\nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[\text{old}]}) + \lambda \boldsymbol{\theta}^{[\text{old}]} \right) \\ &= \boldsymbol{\theta}^{[\text{old}]}(1 - \alpha\lambda) - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[\text{old}]}). \end{aligned}$$

The term $\lambda \boldsymbol{\theta}^{[old]}$ causes the parameter (**weight**) to **decay** in proportion to its size. This is a very well-known technique in deep learning - and simply *L2* regularization in disguise (for gradient descent).

WEIGHT DECAY VS. L2 REGULARIZATION / 2

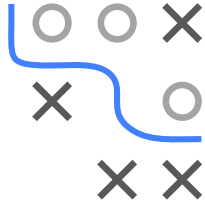


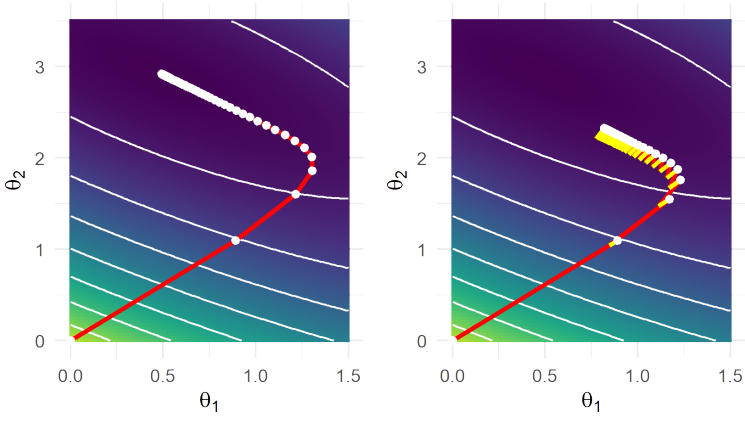
Caveat: Equivalence of weight decay and *L2* only holds for (S)GD!

- ▸ Hanson and Pratt 1988 originally define WD "decoupled" from gradient-updates $\alpha \nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[\text{old}]})$ as $\boldsymbol{\theta}^{[\text{new}]} = \boldsymbol{\theta}^{[\text{old}]}(1 - \lambda') - \alpha \nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}^{[\text{old}]})$
- This is equivalent to modern WD/*L2* (last slide) using reparameterization $\lambda' = \alpha\lambda$
- Consequence: if there is optimal λ', then optimal *L2* penalty is tightly coupled to α as $\lambda = \lambda'/\alpha$ (and vice versa)
- ▸ Loshchilov and Hutter 2019 show no equivalence of *L2* and WD possible for adaptive methods like Adam (Prop. 2)
- In many cases where SGD+*L2* works well, Adam+*L2* underperforms due to non-equivalence with WD
- They propose a variant of Adam decoupling WD from gradient updates (AdamW), increasing performance over Adam+*L2*

WEIGHT DECAY VS. L2 REGULARIZATION / 3

When we use weight decay, we follow the steepest slope of $\mathcal{R}_{\text{emp}}$ as for gradient descent, but in every step, we are pulled back to the origin.

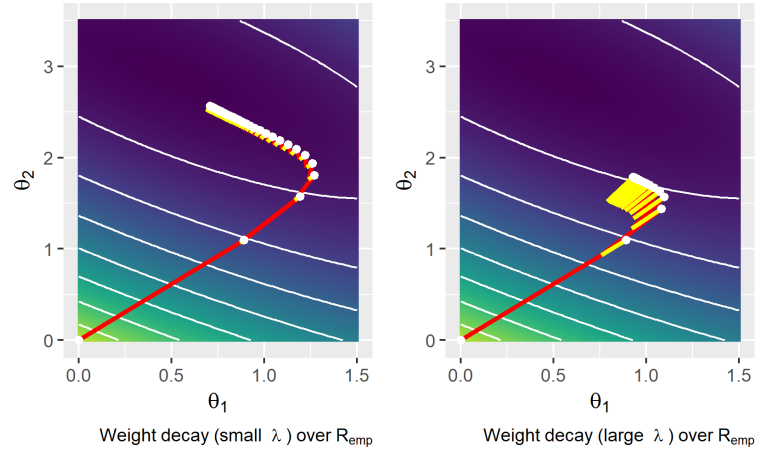




Gradient descent over R_{emp} Weight decay over R_{emp}

WEIGHT DECAY VS. L2 REGULARIZATION / 4

How strongly we are pulled back to the origin (for a fixed stepsize α) depends only on λ as long as the procedure converges:



Weight decay (small λ) over R_{emp} | Weight decay (large λ) over R_{emp}